Are Passages Enough? The MIRACLE Team Participation in QA@CLEF2009

María Teresa Vicente-Díez, César de Pablo-Sánchez, Paloma Martínez, Julián Moreno Schneider, and Marta Garrote Salazar

Universidad Carlos III de Madrid, Avda. Universidad, 30, 28911 Leganés, Madrid, Spain
{tvicente,cdepablo,pmf,jmschnei,mgarrote}@inf.uc3m.es

Abstract: This paper summarizes the participation of the MIRACLE team in the Multilingual Question Answering Track at CLEF 2009. In this campaign, we took part in the monolingual Spanish task at ResPubliQA and submitted two runs. We have adapted our QA system to the new JRC-Acquis collection and the legal domain. We tested the use of answer filtering and ranking techniques against a baseline system using passage retrieval with no success. The run using question analysis and passage retrieval obtained a global accuracy of 0.33, while the addition of an answer filtering resulted in 0.29. We provide an analy-sis of the results for different questions types to investigate why it is difficult to leverage previous QA techniques. Another task of our work has been the appli-cation of temporal management to QA. Finally we include some discussion of the problems found with the new collection and the complexities of the domain.

1 Introduction

We describe the MIRACLE team participation in the ResPubliQA exercise in the Multilingual Question Answering Track at CLEF 2009. The MIRACLE team is a consortium formed by three universities from Madrid, (Universidad Politécnica de Madrid, Universidad Autónoma de Madrid and Universidad Carlos III de Madrid) and DAEDALUS, a small and medium size enterprise (SME). We submitted two runs for the Spanish monolingual subtask which summarize our attempts to adapt our QA system to the new requirements of the task.

The change in the application domain has been triggered by the use of the JRC-Acquis document collection [1], which is formed by European legislation translated in several EU languages. This fact raises the problem of dealing with legal language, which includes technical terminology and shows a more complex syntactic structure than news or academic language used in EFE and Wikipedia collections. Moreover, new information needs require the inclusion of questions asking for objectives, motivations, procedures, etc. in addition to the traditional factual and definition questions. The new types of questions often required longer answers and, therefore, the response of the system has been fixed again at the paragraph level. Nevertheless, it should be possible to take advantage of answer selection techniques developed in previous campaigns. This has been in fact one of the hypothesis we wanted to test. Unfortunately,

our experiments in this line have not been successful and we have not found configurations that performed substantially better than our baseline. Another aspect of our work has focused on the use of temporal information in the process of QA. We report the results for different indexing configurations. Finally, a global objective was to enhance the capabilities of the QA system and advance towards an architecture that allows domain adaptation and multilingual processing.

The paper is structured as follows: section 2 describes the system architecture with special attention paid to the novelties introduced this year; section 3 presents the submitted runs and the analysis of the results. Finally, conclusions and future work are shown in section 4.

2 System Description

The system architecture is based on the approach taken by the MIRACLE QA system participating in CLEF 2008 [2] and consists in a pipeline which analyzes questions, retrieves documents and performs answer extraction based on linguistic and semantic information. A rule engine has been used in the Question Classification, Answer Filter, Timex Analyzer and Topic Detection modules. The left part of the rules are patterns that can refer to lexical, syntactic and/or semantic elements, whereas the right part are actions that add annotations like question types, entity classes or time normalizations. Figure 1 shows the architectural schema.

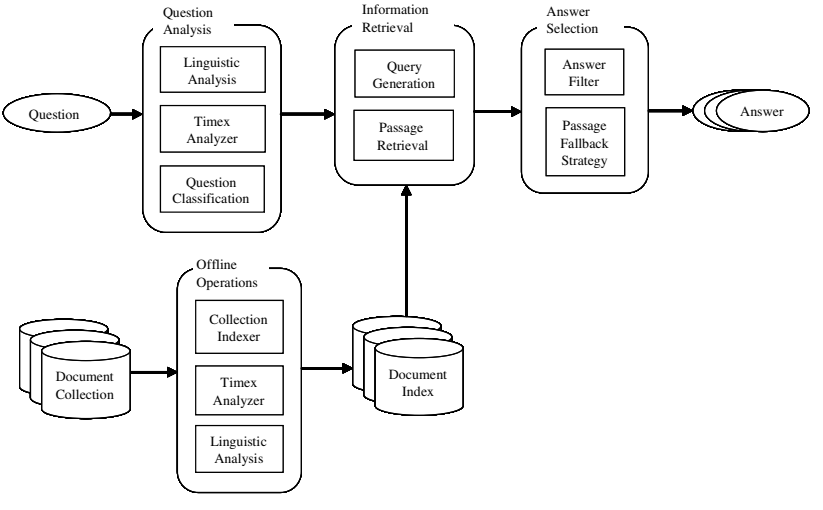

Fig. 1. MIRACLE 2009 system architecture

Some new modules have been included to carry out new experiments while others have been modified, extended or reorganized. The main changes are the following:

- Adding parsers for the new collections and supporting the indexing of passages.
- The evaluation procedure was modified to work with passages and a fallback strategy for passages was also included.
- New rules have been developed for Question Analysis and Answer Selection for the legal domain using the development set.
- Query Generation has been adapted to the domain, removing old heuristics.
- Temporal Management was added and integrated into indexing routines.
- New functionality for mining acronyms and adding them to Query Generation.
- The Ranking module was redesigned for modularity.

2.1 Indexes

Due to the change in the document collection, all IR indexes have been newly created using Lucene [3]. To accomplish the task of storing the relevant information as appropriately as needed, we have designed two different indexing units: *Document*, where all the information related to title, note and the text of the file is stored; and *Paragraph*, which stores each paragraph, title and the notes in a different unit. Lucene uses a length document normalization term in the retrieval score which was arguably of no help in the case of paragraph scoring because paragraphs are expected to have more uniform lengths. Both types of indexes, with and without length normalization, were tested.

In all our experiments the paragraph or passage index worked better than the document index. Besides, we also created different index types regarding the analysis, characterized by the linguistic analyzer used in each case: *Simple Index*, where the text analyzer used is a simple analyzer adapted for Spanish. It makes grammar based parsing, stems words using a snowball-generated stemmer, removes stop words, replaces accented characters and converts text into lower case. *Temporal Index*, which adds recognition and normalization of time expressions. These time expressions are normalized and included in the index.

2.2 Temporal Management

Some authors have defined the temporal question answering (TQA) as the specialization of the QA task in which questions denote temporality [4], as well as a means for providing short and focused answers to temporal information needs [5]. Previous work has already faced up to this problem [6], [7]. Temporal questions can be classified into 2 main categories according to the role of temporality in their resolution: *Temporally Restricted* (TR) questions are those containing some time restriction: *"What resolution was adopted by the Council on 10 October 1994?"*; and *Questions with a Timex Answer* (TA) are those whose target is a temporal expression or a date: *"When does the marketing year for cereals begin?"*

In this campaign, temporal management preserves the approach taken by our previous system [2]. This decision is based on later complementary work that was made in order to evaluate the QA system performance versus a baseline system without

temporal management capabilities [8]. The experiments showed that additional temporal information management can benefit the results.

Several adjustments were made in the temporal expressions recognition, resolution and normalization components to enhance their coverage on the new collections. The date of creation of each document is adopted as the reference date, needed to resolve the relative expressions that the collection could contain (for instance: "*yesterday",* or *"last week*"). This type of expressions need another point in time to be properly resolved, that is, to deduce their semantics. This point of reference could be a date taken from the context of the document but a simpler approach is to consider the date in which contents were created. In JRC-Acquis documents this information is provided by the "*date.created*" attribute. During question analysis process, queries, including those with temporal features, are classified, distinguishing between TR and TA queries. If a TA query is detected, it determines the granularity of the expected answer (complete date, only year, month, etc.). The answer selector is involved in two directions: in the case of TA queries, the module must favour a temporal answer, whereas if it manages TR queries, it applies extraction rules based on the temporal inference mechanism and demotes the candidates not fulfilling the temporal restrictions.

As a novelty, this year we have created more sophisticated indexes according to the paragraph retrieval approach of the competition. In some configurations, the normalized resolution of temporal expressions is included in the index instead of the expression itself [9]. The main objective is to assess the behavior of the QA system using different index configurations, focusing on the temporal queries.

2.3 Acronym Mining

Due to the nature of the collection, a large number of questions were expected to be expansion of acronyms, especially about organizations. On the other hand, the recall of the information retrieval step could be improved by including the acronym and their expansion in the query.

We implemented a simple offline procedure to mine acronyms by scanning the collection and searching for a pattern which introduces a new entity and provides their acronym between parentheses. Then, results are filtered in order to increase their precision. First, only those associations that occur at least twice in the corpus are considered. As parentheses often convey other relations like persons and their country of origin, another filter removed countries (*Spain*) and their acronyms (*ES*) from the list. Finally, some few frequent mistakes were manually removed and acronyms with more than one expansion were also checked. We indexed the acronyms and their expansions separately to be able to search by acronym or by expansion.

The acronym index is used in two different places in the QA system: during Query Generation, where it analyzes the question and adds search terms to the query; and in Answer Filtering, where it analyzes the text extracted from the paragraph to determine if that paragraph contains the acronym (or the expansion) and if so, identifies the paragraph as correct answer.

2.4 Answer Filter and Passage Fallback Strategy

This module, previously called Answer Extractor, processes the result list from the information retrieval module and selected chunks to form a possible candidate answer. In this campaign, the answer must be the complete text of a paragraph and, therefore, this year the module works as a filter which removes passages with no answers. The system has been adapted and new rules to detect acronyms, definitions, as expressed in the new corpora, and temporal questions have been developed.

The possibility of getting no answer from the Answer Filter led to the development of a module that simply creates answers from the retrieved documents. This module is called Passage Fallback Strategy. It takes the documents returned by the information retrieval module and generates an answer from every document.

2.5 Evaluation Module

Evaluation is a paramount part of the development process of the QA system. In order to develop and test the system, the English development test provided by CLEF organizers was translated to Spanish and a small gold-standard with answers was developed. Mean Reciprocal Rank (MRR) and Confidence Weighted Score (CWS) were consistently used to compare the output of the different configurations with the development gold standard. The output of different executions were manually inspected to complete the gold standard and to detect integration problems.

3 Experiments and Results

We submitted two runs for the monolingual Spanish task. They correspond to the configurations of the system that yielded best results during our development using the translated question set.

The first configuration consisted on a version of the system that includes modules for Question Analysis and Information Retrieval together with a number of Offline Operations that perform the linguistic analysis of the collection and originate the indexes. Moreover the management of time expressions (Timex Analysis) was eliminated both in the collection and in the query processing looking for avoiding ambiguity in the semantics of numerical expressions. The second configuration was based on the addition of an Answer Selection strategy to the first design (Figure 1).

We called this runs *mira091eses* and *mira092eses*, each one corresponding to one of the previous configurations as follows:

- Baseline (BL): *mira091eses*. The system is based on passage retrieval using the Simple Index. Question Analysis is performed to generate queries and the acronym expansion is used.
- Baseline + Answer Filter (BL+AF): *mira092eses*. The Answer Filter and the Passage Fallback Strategy modules are added after the previous passage retrieval.

A number of additional configurations were also tested, but no improvements over the baseline were found. In fact, most of the additions seemed to produce worse results. We considered different functions for answer and passage ranking. Different passage

length normalization strategies were also applied to the indexes. Finally, a great deal of effort was devoted to the management of temporal expressions; more detailed experiments are presented below.

Evaluation figures are detailed in Table 1. Answer accuracy (*Acc*) has been calculated as the ratio of questions correctly answered (*Right*) to the total number of questions. Only the first candidate answer is considered, rejecting the rest of possibilities.

Table 1. Results for submitted runs

Name	Rigth	Wrong	Unansw. Right Candidate	Unansw. Wrong Candidate	Unansw. Empty Candidate	Acc.	Correctly discarded	c@1 measure
mira091eses	161	339	0	0	0	**0.32**	0	0.32
mira092eses	147	352	0	0	1	**0.29**	0	0.29

The results on the CLEF09 test set show similar conclusions to those obtained during our development process: the baseline system using passage retrieval is hard to beat; our second run provides lower accuracy. As in the case of our development experiments, there are changes for individual answers of certain questions, but the overall effect is not positive.

We have decided to carry a class based analysis in order to understand the causes behind our unfruitful efforts. We have manually annotated the test set and grouped questions into 6 main types (see Table 2). Contrary to our expectations, the performance of the second submitted run is also worse for the factual and definition questions. As we had considered these questions types in previous evaluations, we expected to have better coverage. Similar behavior has been observed across answer types for factual questions, being the class of temporal questions the only where a more complex configuration really improves.

Our analysis of the errors show that further work is needed to be able to cope with the complexities of the domain. For example, questions are in general more complex and include domain specific terminology that our question analysis rules do not handle correctly. The process of finding the focus of the question, which is crucial for question classification, is specially error prone. Answer Selection needs also further adaptation to the domain for factual questions as the typology of Named Entities (NE) and generalized NE has not wide coverage. On the other hand, being the first time that the legal domain was used, there was not any previous knowledge about how good would be the performance using existing rules of the system in a new context, without a gold standard to suggest some tuning actions.

Problems with definitions are rooted more deeply and probably require the use of different specialized retrieval strategies. This year evidence along with previous experiments seems to support that definitions depend deeply on the stylistics of the domain. Finally, new question types would require further study of techniques that help to improve the classification of passages as bearing procedures, objectives, etc.

Table 2. An analysis of runs by question type

Question Type	TOTAL (test set)	mira091eses Right	mira091eses Acc	mira092eses Right	mira092eses Acc
FACTUAL	123	54	0.44	48	0.39
PROCEDURE	76	22	0.28	15	0.20
CAUSE	102	43	0.42	44	0.43
REQUIREMENT	16	5	0.31	5	0.31
DEFINITION	106	16	0.16	12	0.11
OBJECTIVE	77	21	0.27	23	0.30
ALL	500	161	0.32	147	0.29

Evaluation of Temporal Questions

We extracted the temporal questions from the whole corpus: 46 out of 500 queries denote temporal information, which means a 9.20% over the total. 24 of them are TR questions, whereas TA queries are 22 (4.80% and 4.40% out of the total, respectively). This subset has been studied, evaluating the correctness of the answers by two different configurations of the system. The results are presented in Table 3.

Table 3. Results for temporal questions in the submitted runs and additional configurations

Name	Temporal Questions (TR + TA)	Temporally Restricted (TR)	Timex Answer (TA)
BL (mira091eses)	0.43	0.42	0.45
BL-AF (mira092eses)	0.48	0.37	0.59
DA-BL (non-submitted configuration 1)	0.28	0.21	0.36
DA-BL-AF (non-submitted configuration 2)	0.37	0.21	0.54

Better figures are obtained by the set of TQ in both runs. There is no significant difference between TA and TR queries in the first run, while in the second one they achieve a difference of 22%. In our opinion, the second configuration enhances precision for TA queries, whereas for TR queries, temporal restrictions introduce noise that the system is not able to solve.

Non-submitted runs present similar configurations to the submitted ones, but adopting a different index generation and question analysis strategies. The approach consisted of the inclusion of normalized temporal expressions into the index, as well as into the question analysis process, aiming to increase recall. We tested the performance over the total corpus, but worse results were achieved even if the study is restricted to temporal questions. Results show no improvement regarding the submitted runs. Performance difference between TA and TR queries remains stable, since the system has a better response to questions without temporal restrictions. Once the results were analyzed, we consider incorrect our initial assumption of extracting the reference date from the "*date.created*" attribute of the documents. This hypothesis could be partially the cause of erroneously resolving almost all relative dates. This is due to the fact that we assumed that this attribute was the date of creation of the document, whereas actually it refers to the date of publication of the collection, without providing any significant context information. Besides, lost of accuracy can be due to the lack of a more sophisticated inference mechanism at the time of retrieval, capable of reasoning with different granularities of normalized dates.

4 Conclusion and Future Work

From our point of view, the new ResPubliQA exercise is a challenge for QA systems in two main senses: linguistic domain adaptation and multilingualism. This year our efforts have focused on the first problem, adapting the previous system to the new collection. However, our experiments show that a system mainly based on passage retrieval performs quite well. Baseline passage retrieval results provided by the organizers [10] also support this argument. We are carrying out further experiments to find how answer selection could help for ResPubliQA questions, as well as the differences between passage retrieval alternatives. Regarding our task on temporal reasoning applied to QA, we will explore how question temporal constraints can be integrated at other steps in the process. We expect to compare the effectiveness of temporal reasoning as constraints for filtering answers and for the purpose of re-ranking. Finally, further work in the general architecture of the QA system is planned regarding three areas: separation of domain knowledge from general techniques, adding different languages to the system and effective evaluation.

Acknowledgements. This work has been partially supported by the Research Network MAVIR (S-0505/TIC/000267) and by the project BRAVO (TIN2007-67407-C3-01).

References

1. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, Italy (2006)
2. Martínez-González, A., de Pablo-Sánchez, C., Polo-Bayo, C., Vicente-Díez, M.T., Martinez-Fernández, P., Martínez-Fernández, J.L.: The MIRACLE Team at the CLEF 2008 Multilingual Question Answering Track. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 409–420. Springer, Heidelberg (2009)
3. Apache Lucene project. The Apache Software Foundation, `http://lucene.apache.org`
4. Saquete, E.: Resolución de Información Temporal y su Aplicación a la Búsqueda de Respuestas. Thesis in Computer Science, Universidad de Alicante (2005)
5. De Rijke, et al.: Inference for temporal question answering Project (2004-2007)
6. Clark, C., Moldovan, D.: Temporally Relevant Answer Selection. In: Proceedings of the 2005 International Conference on Intelligence Analysis (2005)
7. Saquete, E., Martínez-Barco, P., Muñoz, R., Vicedo, J.: Splitting Complex Temporal Questions for Question Answering Systems. In: Proceedings of the ACL 2004 Conference, Barcelona, Spain (2004)
8. Vicente-Díez, M.T., y Martínez, P.: Aplicación de técnicas de extracción de información temporal a los sistemas de búsqueda de respuestas. Procesamiento del lenguaje natural (42), 25–30 (2009)
9. Vicente-Díez, M.T., Martínez, P.: Temporal Semantics Extraction for Improving Web Search. 8th International Workshop on Web Semantics (WebS 2009). In: Tajoa, A.M., Wagner, R.R. (eds.) Proceedings of the 20th International Workshop on Database and Expert Systems Applications, DEXA 2009, pp. 69–73. IEEE Press, Los Alamitos (2009)
10. Pérez, J., Garrido, G., Rodrigo, A., Araujo, L., Peñas, A.: Information Retrieval Baselines for the ResPubliQA task. In: CLEF 2009 Working Notes (2009)